UCLA Anderson

SCHOOL of MANAGEMENT

MGMTMSA 405: Data Management - MSBA 2024

Final Report

Project 4 - Build Data Warehouse for Crime Data

Group 19

Lodha, Chitransh	306308898
Yu, Irvy	306310684
Zhang, Zhi	306311909
Klappenbach Zucchi, Facundo	306316186
Cao, Cocoon	406082115

Table of Content

Executive Summary

A concise data warehouse was meticulously developed for the crime data from Los Angeles from 2020 to the most recent date. A thorough analysis was conducted, culminating in the seamless integration of the data within Snowflake, which facilitated the creation of eight structured fact and dimension tables. These tables served as a robust foundation for executing intricate SQL queries essential for deriving five critical performance indicators (KPIs). Subsequently, these KPIs were brought to life through dynamic and interactive visualizations within the Tableau dashboards, offering clear and actionable insights.

The data revealed a general uptick in crime rates from 2020 to 2023, with a 5.22% surge in 2022. Seasonality played a role, as crime rates spiked in the third quarter each year, with the first quarter consistently showing a reduction. Specific areas like the Central region saw heightened crime rates, suggesting targeted areas for policing efforts. In contrast, areas such as Hollenbeck and Foothill maintained stable crime rates over the four-year period. Noteworthy declines in crime rates in Central, Rampart, and Mission by 6.2%, 5.0%, and 3.9%, respectively, indicate effective law enforcement interventions.

In response, strategic recommendations include intensifying crime reduction initiatives in areas with rising crime rates and enhancing surveillance and resource allocation in areas with early signs of increased crime rates. Additionally, the insights call for replicating successful strategies from regions with declining crime rates and adapting policing and community programs to address seasonal crime spikes.

Our findings also encompass crime types, with assault-related crimes dominating. There was a significant 53% increase in burglary cases from 2022 to 2023, warranting further investigation. The report also highlights time-based patterns in criminal activity, particularly in the afternoon and evening hours, guiding recommendations for law enforcement presence during peak crime hours.

In conclusion, the project report and Tableau dashboard serve as comprehensive tools for analyzing crime trends, informing law enforcement strategies, and guiding resource allocation to enhance public safety in Los Angeles. The successful completion of this project required and demonstrated strong data analytics, database management, and data visualization skills, along with effective communication and teamwork.

Project Statement

The dataset represents crime incidents occurring in Los Angeles from 2020 onwards. The information is derived from original crime reports written on paper, which may result in occasional inaccuracies. Instances where location details are absent are marked as (0°, 0°), while address information is rounded to the nearest hundred block to safeguard privacy. Despite potential discrepancies, the data aims to maintain the accuracy of the database and thus, shed light on crime incidents in LA.

Implemented to enhance the overall quality of crime data gathered by law enforcement, the National Incident-Based Reporting System (NIBRS), records comprehensive information on individual crime incidents, including details on victims, perpetrators, relationships between them, arrestees, and property involved, which leads to the dataset that we are analyzing.

This data offers deep insights by capturing circumstances and context such as location, time, and incident clearance status. Recognized as a crucial tool by law enforcement organizations, the FBI prioritizes nationwide adoption of NIBRS to furnish more informative statistics for productive discourse, strategic planning, and informed policing.

The project goals revolve around effective collaboration, application of knowledge, and gaining experience in a real-world analysis example from scratch.

Firstly, teamwork was emphasized, participation from all members was expected, and individual accountability was enforced to ensure an equal contribution from everyone. Secondly, and most importantly, this project aims to leverage the learnings from the data management class. This entails applying theoretical concepts and analytical skills acquired throughout the program to analyze real-world data effectively. As an example, the next concepts learned in the data management course were applied:

1. A clear understanding of cloud computing and its benefits to use in this project (Snowflake usage)
2. Database Design Consideration
3. Data Integration benefits, Extract Transform & Load (ETL), Data Issues
4. Data Transformation
5. Data Governance & Data Dictionary
6. Dimensional Modeling, Star Schema for reporting & analytics
7. Fact & Dimention tables
8. Visualization with Tableau

Lastly, strong communication skills and creativity are essential components of the project. While encouraging creativity in generating ideas, it's crucial to maintain relevance to the project's objectives. The ultimate goal is to produce a professional-quality report, reflective of the standards expected in a corporate environment. This emphasis on professionalism underscores the importance of attention to detail and quality in the final deliverable.

Data Dictionary

For this project, we had two raw datasets extracted from datala.city.org (Los Angeles city public data). The first dataset (CRIME_DATA_LAPD) is transactional data and contains the information of each crime incident. The second dataset has information about the perpetrator's modus operands, which can be connected to CRIME_DATA_LAPD for more information.

After analyzing and understanding the raw data and the use case, we created the star schema (see Dimensional Modeling section) and created new fact and dimension tables to analyze the information in a simpler, organized way:

Table name	Number of columns	Number of rows	Description
CRIME_DATA_LAPD_FIXED	33	910,707	Raw data with minor formatting changes
MO_CODE_DESC	2	777	MO codes description raw data
FACT_CRIMEINCIDENTS	9	910,707	Fact table of incidents with needed keys to join with dimensions
CRIME_DIM	3	139	Crime dimensions table with info about the crime
DATE_DIM	26	15,000	Table with Information about the date when the crime occurred
LOCATION_DIM	10	910,707	Table with information about the location where the incident occurred
STATUS_DIM	3	6	Table with information about the status of the incident. (if the criminal was arrested or not)
TIMEOFDAY_DIM	7	1,440	Table with information about the time of day
VICTIM_DIM	5	910,707	Table with information about the

			demographics of the victim
WEAPON_DIM	3	79	Table with information about the weapon characteristics used by the criminal

Finally, information about the content of each table and description of each column, as well as data type can be found in the Dictionary annexed into the file deliverable.

Dimensional Modeling + ERD

Step1: Define the Business or Event Process

The business process here is the reporting and recording of crime incidents. Each record represents a crime event with details about the crime, victim, location, time, and other attributes.

Step2: Define the Grain

The grain of the fact table is a single crime incident. This granularity allows for the recording of unique instances of crimes, which is necessary for detailed analysis.

Step3: Identify the Dimensions

VICTIM_DIM: Chosen because victim details are critical for demographic analysis of crime impact.

WEAPON_DIM: Relevant for understanding the types of weapons used in crimes.

LOCATION_DIM: Essential for spatial analysis of crime patterns and hotspots.

CRIME_DIM: Includes types of crimes, necessary for categorizing and analyzing crime statistics.

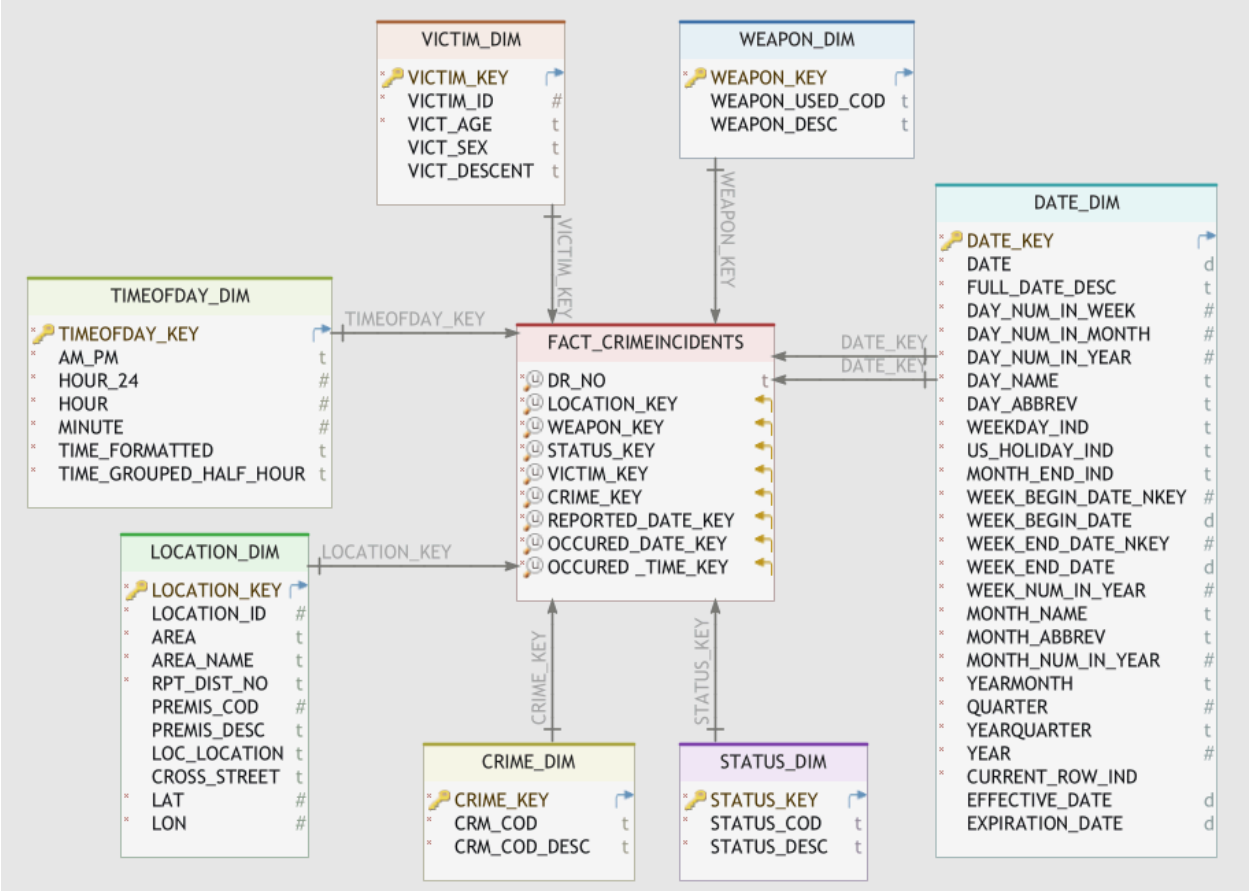
TIMEOFDAY_DIM: Important for analyzing crime trends over the course of a day.

STATUS_DIM: Could be used to track the status of the crime investigation (solved, unsolved, etc.).

DATE_DIM: Key for time-based analysis and trend observation over dates, months, and years.

Step4: Identify the Fact

We created our fact table: FACT_CRIMEINCIDENTS, which is a confluence of measurable attributes and keys that link to our dimension tables. This central fact table is designed to facilitate complex queries, supporting the exploration of trends, patterns, and correlations.



Exhibition 2

Exhibition 2 presents the ER Diagram of our structured crime database, with the FACT_CRIMEINCIDENTS table positioned at the center of our star schema. This table is a confluence of key measurable attributes and foreign keys that reference our dimension tables—each serving as a repository of rich descriptors that illuminate various facets of the data.

Each dimension table encapsulates a specific aspect of the data, fostering a multidimensional analysis landscape:

VICTIM_DIM delves into the human element, painting a detailed picture of those affected by crime, and serves as a pivotal tool for socio-demographic analyses.

WEAPON_DIM catalogs the instrumentalities of crime, enabling pattern recognition in weapon usage and potentially aiding in preventative measures.

LOCATION_DIM offers geographical intelligence, vital for pinpointing hotspots and deploying resources effectively.

CRIME_DIM classifies each incident, contributing to trend analyses and strategy formulation for crime reduction.

TIMEOFDAY_DIM and DATE_DIM dissect the temporal data, offering insights into crime cyclicity and temporal patterns that could influence policing strategies.

STATUS_DIM tracks the progress of crime resolution, an essential measure of operational effectiveness and judicial performance.

Leveraging these dimensions, we can easily make a variety of complex queries. The schema facilitates swift explorations into trends, patterns, and correlations, ultimately enabling stakeholders to derive actionable insights and make informed decisions.

Data Transformation:

1. CRIME_DIM

The crime_dim exhibits the crime description for each crime code, crm_cd. To avoid repetition, we used the SELECT DISTINCT operator to select unique crime codes and their corresponding description from the crime_data_lapd_fixed dataset. We inserted an auto-incremented crime_key as the primary key for each record.

2. STATUS_DIM

Similar to the crime_dim, each record in the status_dim provides information about the status code and the description. Therefore, in a similar procedure, we selected unique status codes and the corresponding description from the dataset and used sequence to generate auto-incremented keys.

3. WEAPON_DIM

Weapon_dim also stores the record for weapon_used_cod and the description weapon_desc. Similarly, we used SELECT DISTINCT to select a unique weapon code from the dataset. The primary key weapon_key was also generated with an auto-incremented sequence.

4. LOCATION_DIM

Compared to other dimensions, location_dim contains more information that details the specific location for each crime record. Since the premise description, street information, and specific

location are unique for each crime case and could be helpful for further analysis, we decided to populate every row of location information from the original dataset into the dimension table. As location does not have a unique key, like weapon_used_cod to the weapon, it could be challenging to connect location dim with the fact table, and relevant information would be difficult to retrieve. Therefore, we decided to create a new table LOCATION to design a new key, LOCATION_ID, to connect location information with each crime record. In the LOCATION table, the new key LOCATION_ID is created through auto-incrementation, and the DR_NO was also included as a foreign key to connect with the crime dataset. In this way, the location information could be smoothly populated in the dimension.

5. VICTIM_DIM

The victim information also varies across different crime records. However, it also lacked a unique key to connect the dimension with other variables for the crime records. Therefore, we duplicated the procedure for building location_dim and created a separate VICTIM table to store the connection of VICTIM_ID and DR_NO. With the unique VICTIM_ID key, we were able to populate data into the dimension.

6. DATE_DIM

In order to explore seasonality patterns, we included multiple variables for different granularity, such as the day in the week, month, year, and holiday periods. We populated data from the GENERATOR that was set to generate 20 years and used the CASE WHEN clause to determine the specific variable for each date. The primary key DATE_KEY is defined as the “yyyymmdd” format for each date for easier reference.

7. TIMEOFDAY_DIM

The timeofday_dim is also populated with generated data. In order to record time data at the minute level, we used a Python query to generate the INSERT query for each row from 00:00:00 to 23:59:00. We also created a TIME_GROUPED_HALF_HOUR variable for future exploration.

8. FACT_CRIMEINCIDENTS

The FACT_CRIMEINCIDENTS table connects all the dimension tables with the dimension keys. In this fact table, dr_no is the primary key that represents each crime record in the original dataset. Using the crm_cod, weapon_used_cod, and status_cod from the dataset, we were able to connect the CRIME_DIM, WEAPON_DIM, and STATUS_DIM to the fact table. As we created separate tables for LOCATION and VICTIM, we used the location_id and victim_id to bridge the dimension tables with the primary key dr_no, thus connecting them to the fact table. For the date information, we would like to record the report date, occurred date, and occurred time in the fact table. Therefore, we joined these tables by setting date_rptd, date_occ, and time_occ from the original dataset equal to date and time_formatted from the dimension tables. Notably, the date_dim table was joined two times with one recording the reported day information and the other showing the occurred date information.

Produce KPIs using Fact and Dim:

KPI 1: Crimes frequency by area

This KPI measures the total number of crimes in each area, for a given time period. We calculated the KPI by joining the location_dim table with the fact_crimeincidents table using the same location_key. After retrieving the information of the location, we could measure the frequency of crime in each area by grouping the data with area_name and calculating the total count of crime_key in each group.

	AREA	YEAR	FREQUENCY
1	77th Street	2024	2260
2	77th Street	2023	14296
3	77th Street	2022	14317
4	77th Street	2021	13034
5	77th Street	2020	12980
6	Central	2024	2782
7	Central	2023	15312
8	Central	2022	17211
9	Central	2021	14031
10	Central	2020	12735

KPI 2: Crime Frequency by Crime Type

This KPI measures the total number of crimes for each crime type. We also incorporated the option to review the crime frequency of each crime type during different time periods or different areas to explore potential patterns. Similar to the first KPI, we joined the fact_crimeincidents table with crime_dim table to examine the frequency of crimes that occurred for a specific crime. In this way, we grouped the data with different crime types and calculated the count of crime_key. Notably, we selected crm_cod_desc, the description of each crime type, instead of the crime code for better user visibility.

	CRIME_TYPE	YEAR	FREQUENCY
1	ARSON	2024	70
2	ARSON	2023	573
3	ARSON	2022	531
4	ARSON	2021	627
5	ARSON	2020	665
6	ASSAULT WITH DEADLY WEAPON ON POLICE OFFICER	2024	26
7	ASSAULT WITH DEADLY WEAPON ON POLICE OFFICER	2023	179
8	ASSAULT WITH DEADLY WEAPON ON POLICE OFFICER	2022	227
9	ASSAULT WITH DEADLY WEAPON ON POLICE OFFICER	2021	217
10	ASSAULT WITH DEADLY WEAPON ON POLICE OFFICER	2020	397

KPI 3: Crime frequency by time of day

This KPI measures the number of crimes occurring in the morning, afternoon/evening, or at night. We joined the timeofday_dim and fact_crimeincidents tables to cross-examine the crime

information with detailed time variables. To measure the frequency in each time of day, we used CASE WHEN clause to create groups for these time periods based on hour in the day, hour_24. Based on the sunrise time period, we defined anytime from 4 am to 12 pm as morning, 12 pm to 8 pm as afternoon/evening, and 8 pm to the next day 4 am as night. With the definition, we selected the time period and calculated the total count of crime_key under each group.

	YEAR	TIME_OF_DAY	FREQUENCY
1	2020	afternoon/evening	87587
2	2020	morning	56679
3	2020	night	55351
4	2021	afternoon/evening	89640
5	2021	morning	60972
6	2021	night	58984
7	2022	afternoon/evening	95782
8	2022	morning	72363
9	2022	night	66639
10	2023	afternoon/evening	100232

KPI 4: Victim frequency by age range and sex

This KPI measures the number of victims for each group of sex and age groups. To examine the victim data, we joined the victim_dim table with the fact_crimeincidents table through the victim_key. As we would like to explore the pattern of victim frequency in different age ranges and sex, we created 4 groups to segment ages: under 18, 18-24, 25-64, and above 65. With this age division, we would be able to examine the pattern for cohorts of teenagers, young adults, adults, and seniors. Additionally, we removed sex with null value for this KPI for information brevity. The KPI was calculated by counting the total crime_key after grouping by age group and sex.

	AGE_GROUP	SEX	···	FREQUENCY
1	18-24	F		43150
2	18-24	H		15
3	18-24	M		36590
4	18-24	X		6834
5	25-64	F		247839
6	25-64	H		78
7	25-64	M		269128
8	25-64	X		1452
9	A65	F		24114
10	A65	H		6

KPI 5: Weapon frequency by time of day

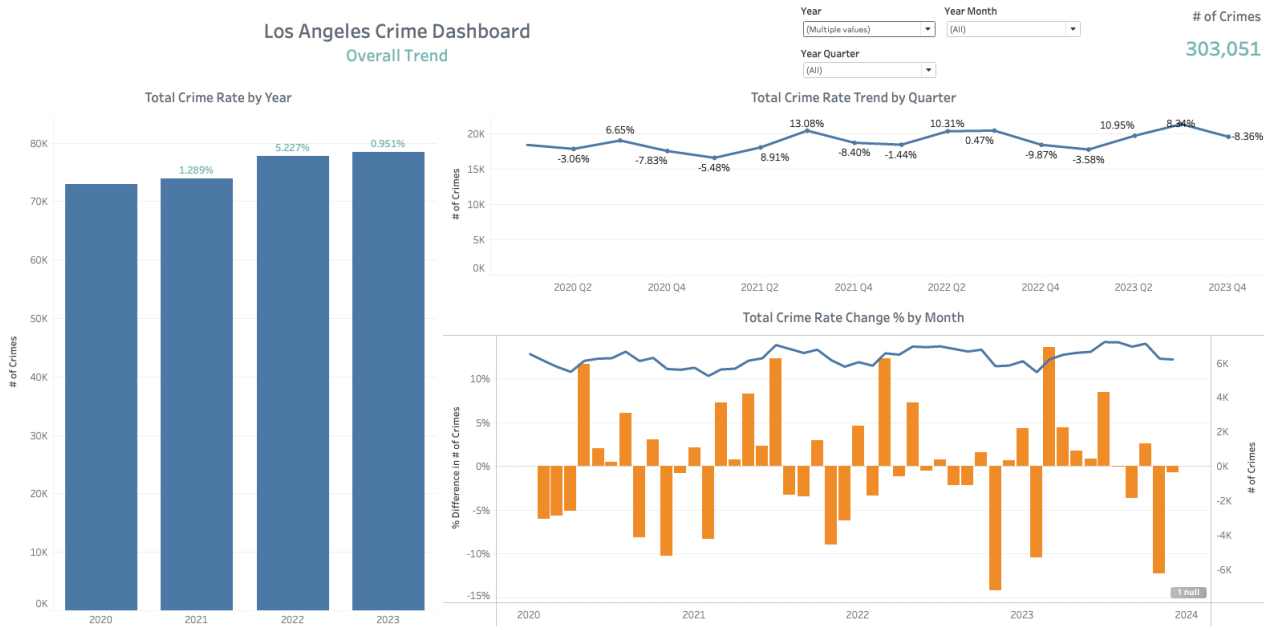
This KPI measures the number of crimes that use a given weapon for each time of day. We joined weapon_dim, date_dim, and fact_crimeincidents tables to select information about the weapon and time of day. Using the insights from KPI 3, we created groups of time based on the hour in the day when the incident occurred. For better visibility of the KPI, we used

weapon_desc, the description of a specific weapon type instead of the code to highlight the weapon frequency.

	TIME_OF_DAY	WEAPON_TYPE	WEAPON_FREQUENCY
8	afternoon/evening	BLUNT INSTRUMENT	559
9	afternoon/evening	BOARD	136
10	afternoon/evening	BOMB THREAT	75
11	afternoon/evening	BOTTLE	940
12	afternoon/evening	BOW AND ARROW	7
13	afternoon/evening	BOWIE KNIFE	20
14	afternoon/evening	BRASS KNUCKLES	103
15	afternoon/evening	CAUSTIC CHEMICAL/POISON	98
16	afternoon/evening	CLEAVER	13
17	afternoon/evening	CLUB/BAT	860

Data Visualization with Important Insights and Recommendations

The total number of crimes reveals a continued increase from 2020 to 2023 in Los Angeles, with a noticeable rise of 5.22% in 2022. The quarterly data uncovers a seasonal pattern where crime peaks in the third quarter of each year, suggesting a possible correlation with seasonal factors or events, while the first quarter consistently records the lowest crime rates, indicating a cyclical lull. Monthly trends display volatility with peaks and troughs. It shows a general increase in crime from March to July. The crime frequency often rises in October, followed by a sharp decline in November each year. This cycle suggests a seasonal pattern with spring and summer months experiencing a surge and early winter seeing a significant reduction in crime rates.



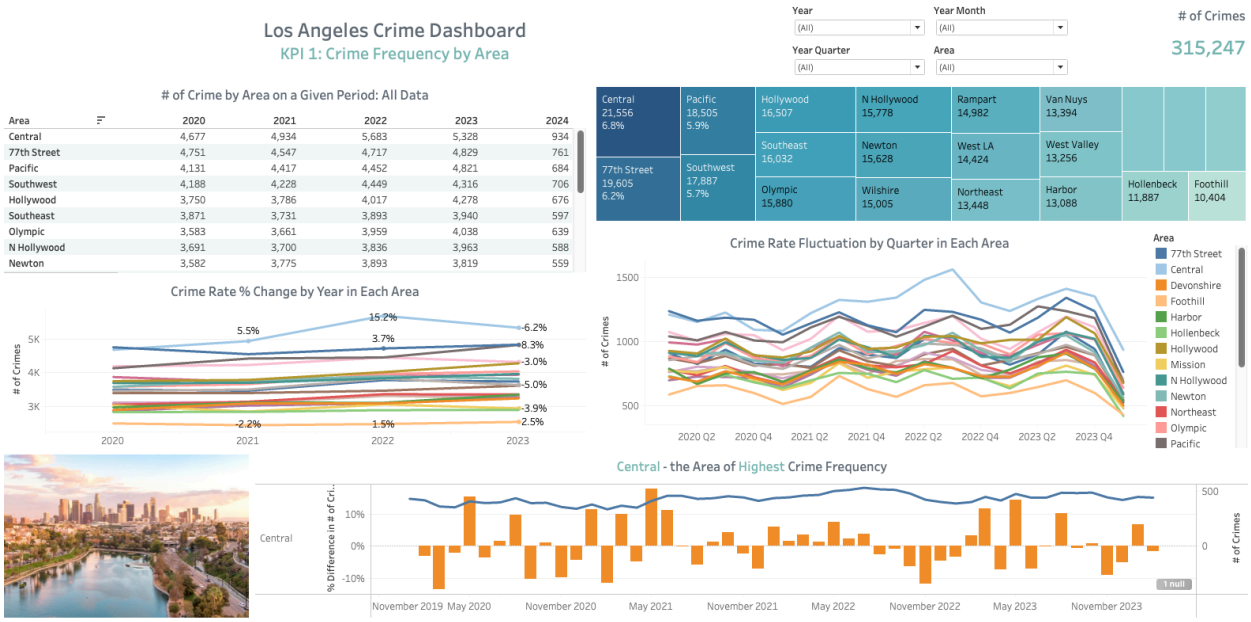
Total # of Crimes in LA from 2020 to the Present

KPI 1: Crimes frequency by area:

The **Central** area emerges as a critical focus, leading with a 6.8% crime rate, trailed by 77th Street and Pacific with 6.2% and 5.9%, respectively. An analysis of annual crime trends reveals that Hollenbeck and Foothill maintain a steady rate of crime incidents, while most areas display a rising trend in crime from 2020 to 2023. Central, Rampart, and Mission are noteworthy for their substantial crime reductions – 6.2%, 5.0%, and 3.9%, respectively, from 2022 to 2023 – hinting at the success of the strategies employed by local law enforcement. Conversely, the Pacific, Harbor, and Hollywood areas have encountered marked increases in crime, suggesting an urgent need for strategic resource allocation and enhanced crime prevention measures. Quarter-over-quarter data indicates that crime rates across all areas demonstrate similar patterns of fluctuation. Central's crime rates show considerable quarterly variance, peaking in Q3 of 2022. May, in particular, marks the sharpest rise in Central, echoing a broader regional pattern.

Recommendations/Actions:

- Prioritize crime reduction efforts in 77th Street, Pacific, and Hollywood, where high crime rates continue to climb.
- Allocate additional monitoring and resources to Central, 77th Street, and Southwest in 2024 as early data points to an upward crime trend.
- Implement proactive policing and community programs during the second and third quarters, focusing on historical peaks, i.e. third quarter, to curb the seasonal rise in crime.
- Study the strategies deployed in Central, Rampart, and Mission that contributed to crime reduction for potential replication in areas where crime is increasing.



# of Crime by Area on a Given Period: All Data

Area	2020	2021	2022	2023	2024
Central	848	714	881	789	889
77th Street	828	713	681	702	725
Pacific	683	640	687	732	651
Southwest	722	622	707	635	677
Hollywood	647	557	621	656	651
Olympic	628	578	636	592	608
Southeast	678	532	568	573	568
Newton	646	568	585	569	523
N Hollywood	604	552	550	570	566

# of Crimes by Area on Jan and Feb in 2024 compared to previous years

KPI 2: Crimes frequency by Crime Type:

The most popular types of crime based on frequency are **Battery - Simple Assault** (23%), Assault with Deadly Weapon, Aggravated Assault (16.4%), and Intimate Partner Simple Assault(14.3%). These three together constitute 54% of the overall crimes in Los Angeles Region.

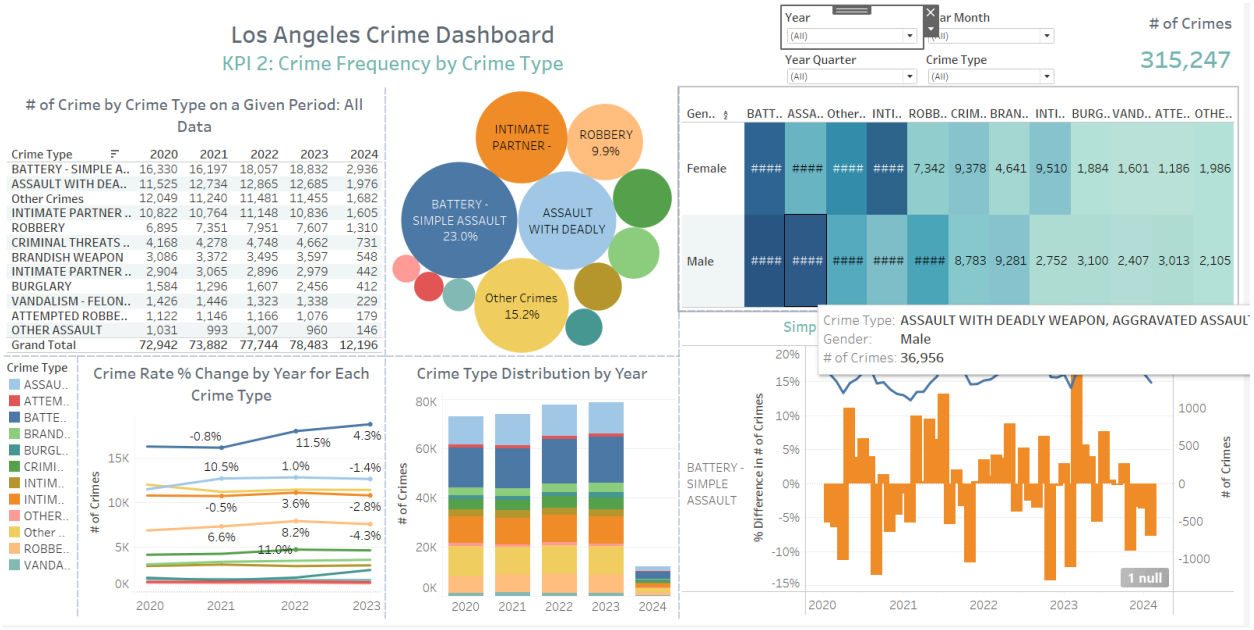
Among Men and Women, we see a similar trend of Battery - Simple Assault, Assault with Deadly Weapon, and Other Crimes as the top 3 crime types. In the last year, there has been a steady rise in cases of Battery - Simple Assault (4.3%), Intimate Partner Aggravated Assault (2.9%), Vandalism (1.1%), and Burglary(53%). The most significant increase is in burglary incidents - a 53% increase from 1,607 cases in 2022 to 2,456 cases in 2023.

All other crime categories have seen a reduction from 2022 to 2023, with robbery showing a 4.3% decline, Deadly Weapon Assault 1.4% decline, Criminal Threats 1.8% decline and other assaults at 4.7%. The most significant decrease was in Robbery cases from 7,951 in 2022 to 7,607 cases in 2023.

For Simple Assaults - the most common type of crime we see a seasonal pattern of reduction in crime frequency during the months of November, December and the peaks during mid-year around the months of July - August. This trend is in line with the overall total crime trend observed.

<u>Recommendations/Actions:</u>

- More analysis for the type of crime committed in Battery - Simple assaults to decode the cause behind the cyclical pattern of crime frequency.
- Investigation for the cause of a significant rise (53%) in burglary cases needs to be done.
- The other categories of crime including simple assault, intimate partner aggravated assault, and vandalism which have seen a steady rise need to be controlled and appropriate action to be taken.



# of Crimes by Crime Type

KPI 3: Crimes frequency by Time of Day:

Based on the data we divided 24 hours into three 8-hour windows Morning, Afternoon/Evening, and Night. The most significant cases were seen in the **Afternoon/Evening between 11 AM - 8**

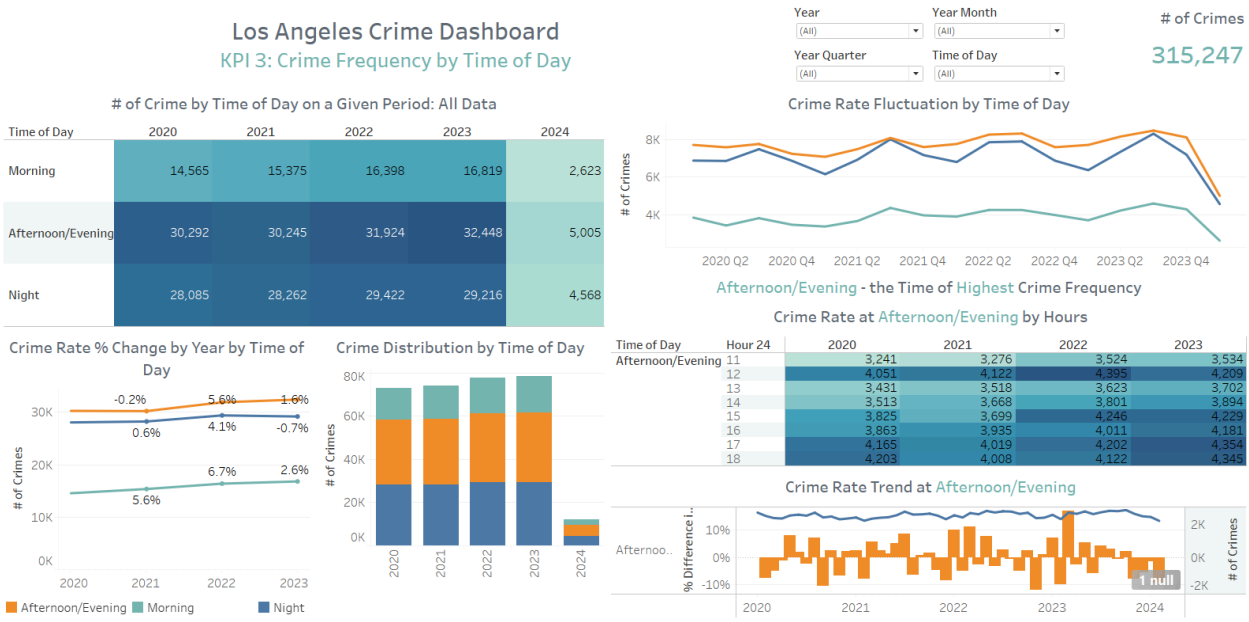
PM with 40% of the crimes in this period followed by Night which sees 37% of the crimes and Morning with 23%.

The number of crimes in Night decreased by 0.7% in 2023 compared to 2022 whereas the is an increasing trend for both Crimes conducted in the morning (2.6%) and during Afternoon/Evening (1.6%). We see a similar seasonal trend with most crimes being reported in Quarter 2 and Quarter 3, while the least incidents were reported in Quarter 1.

The highest number of crimes in the Afternoon/Evening hours is seen between 3 PM - 7 PM and between 12 PM - 1 PM.

<u>Recommendations/Actions:</u>

- The data reveals that most of the crimes happen in the Afternoon/Evening hours and so appropriate security measures need to be enforced to reduce the crime.
- With most crimes reported between 3 PM - 7 PM and between 12 PM-1 PM, more police force can be deployed at public places to enhance security and reinforce preventive measures.



# of Crimes by Time of Day

KPI 4: Victim frequency by age range and sex:

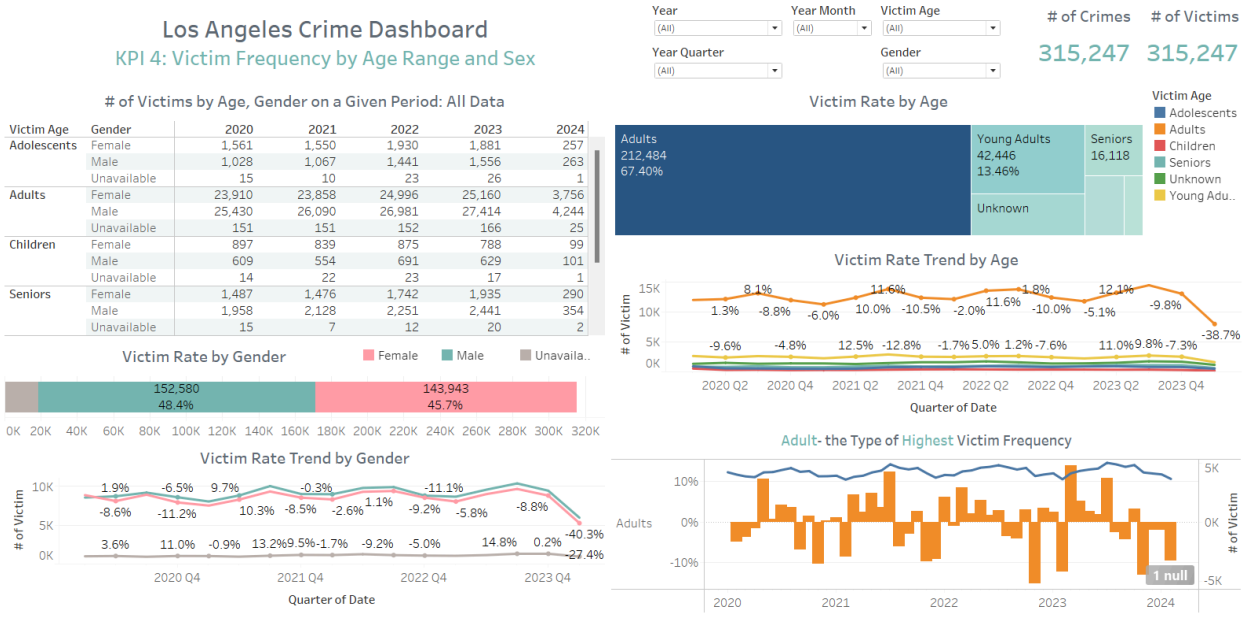
The data reveals that most of the crimes are committed by **Adults** (67.4%) followed by Young Adults (13.5%) and then Seniors (5.115).

In terms of gender distribution, both Males and Females have similar distribution with each Male leading with 48.4% and Females with 45.8%.

We see a reduction in crime rates among Children going down by 16.4% from 2023 Q3 to 2023 Q4. There is also an overall reduction in crime frequency among adolescent-female and Children Females from 1,930 in 2022 to 1,881 in 2023. Whereas the crime frequency among Males in the Adolescent category increased from 1,441 in 2022 to 2023 but reduced from 691 to 629 in the Children age group.

<u>Recommendations/Actions:</u>

- Suitable corrective measures need to be taken to control the rise of crime incidents among both Adult and Senior groups for both the Male and Female categories.
- The crime frequency among Children and Adolescents has seen a downward trend and suitable actions need to be taken to reinforce this trend further.



# of Crimes by Age Range and Sex

KPI 5: Weapon frequency by time of day:

The data reveals that over 60% of the crime incidents happened with **bodily force or without any weapons**, followed by 12% - Unknown Weapons, 8% verbal threats and 7% with a Hand Gun among others.
There is a decrease in year-on-year crime frequency among Hand Gun, Strong Arm, Knife, Semi Automatic pistol crimes.
Whereas, the data shows an increase in crime frequency from verbal threats, blade knives, pepper spray, and unknown category.

<u>Recommendations/Actions:</u>

- Suitable corrective measures need to be taken to control the rise of verbal threats, blade knives, and pepper spray incidents by further analyzing the root cause of these issues.
- There is a large number of crimes without a weapon type identified(12.3%), further analysis needs to be done to understand the cause of these crimes and curtail their spread.

# of Crimes by Weapon Type and Time of Day

Project Challenges

<u>Challenges to upload the data into the snowflake:</u>

The first challenge was uploading the data into Snowflake. The different formatting within the columns was a problem, so we had to account for that before loading it.

- DATE_RPTD and DATE_OCC, some rows had “AM” included in the formatting and some didn’t. To clean this we ran a Python code that only stracted the date.
- TIME_OCC was supposed to include data on the time when the incident occurred in military format. When we checked, there were cases where the military time was only 1, 2, 3 etc. We assumed those numbers meant minutes after 00:00, so we fixed it and modified the column. Also, we created a military time column named TIME_OCC_GROUPED with the time grouped in 30-minute intervals for better interpretation.
- MOCODES is the “Modus Operandi” code of the crime, we notice that all the codes were concatenated in one single row, and it would be hard to join those codes with the

description. Therefore, we created four new columns with the first four codes for simplicity and understanding of the data.
- MO_CODES data was a PDF, so we had to transform that into a CSV file so we could upload it.

After these modifications, we successfully uploaded the CSV data with the formatting changes to Snowflake.

Challenges to create the fact and dimension tables:

Since the original dataset was a single sheet containing all the columns of data, it was also challenging to categorize it into separate tables and determine the keys connecting each dimension. For example, victim and location information is unique in each row of the crime record. However, these two categories do not have a superkey associated with the category other than the primary key dr_no. Thus, if we only include category information in the dimension, there is no connecting key to join the dimension table to the fact table. To solve this dilemma and avoid redundancy, we decided to create new keys for these two categories, victim_id and location_id. To join the new keys to the original dataset, we created the new tables LOCATION and VICTIM to store the relationship to dr_no and the new IDs. We also inserted the IDs into the appropriate dimension tables for future reference. In this way, while populating the fact table with dimension keys, we can directly use the created IDs to join dimensions and datasets, saving processing time and reducing query complexity.

Challenges when we had to analyze the data in Tableau:
- The time of day crime data was given by hours and it was difficult to analyze the data at an hourly level. So, we grouped the hourly time of day data into three buckets of Morning (3AM - 10AM), Afternoon/Evening (11AM - 6PM), and Night (7PM - 2AM) to create categories and better analyze the crime distribution by time of day.
- Similarly, to analyze the crime frequency by age group we grouped age to form buckets of Children, Young Adults, Adults, and Seniors. This helped us to analyze crime trends among various age groups.
- There was a large number of weapon types with less than 100 crime incidents and it was not practical to assess trends for such small categories. So, we only considered the top 13 weapon types and the rest were combined into a larger miscellaneous category. A similar combination was done for analyzing crime frequency by crime type.
- The gender information for about 6% of the columns (18,724 victims) was unavailable in the dataset
- In the crime type field as well, there were about 100 crime types categories with less than 1,000 crime incidents and so we grouped them together as Other Crimes and considered only the Top 11 Crime Types in our dashboard.